HAMILTON, BROOK, SMITH & REYNOLDS, P.C.
530 Virginia Road, P.O. Box 9133
Concord, MA 01742-9133

Telephone: (978) 341-0036

Facsimile: (978) 341-0136

UNOFFICIAL DOCUMENT FOR EXAMINER'S REVIEW

FACSIMILE COVER SHEET

Examiner: Ms. Lynne Black **Group:** 2171

Date: August 5, 2004

Client Code: 2471

Facsimile No.: 703-872-9306

From: David J. Thibodeau, Jr.

Subject: **Paper:** **Method and System for Finding Similar Records in Mixed Free-Text and Structured Data**

Docket No.: 2471.2001-001

Applicants: Eric Bloedorn

Serial No.: 10/091,932

Filing Date: March 6, 2002

Number of pages including this cover sheet: 7

Please confirm receipt of facsimile: Yes x No ___

Comments:

HAMILTON
BROOK
SMITH &
REYNOLDS, P.C.

PATENTS, TRADEMARKS
COPYRIGHTS & LITIGATION

530 VIRGINIA ROAD
P.O. BOX 9133
CONCORD, MA 01742-9133
TEL (978) 341-0036
FAX (978) 341-0136
www.hbsr.com

MUNROE H. HAMILTON
(1909-1984)

DAVID B. BROOK
JAMES M. SMITH
LEO R. REYNOLDS
JOHN L. DUPRE
DAVID J. BRODY
MARY LOU WAKIMURA
ALICE O. CARROLL
N. SCOTT PIERCE
HELEN F. WENDLER
SUSAN G. L. GLOVSKY
DOREEN M. HOGLE
RICHARD W. WAGNER
ROBERT T. CONWAY
RODNEY D. JOHNSON
DAVID J. THIBODEAU, JR.
ANNE J. COLLINS
TIMOTHY J. MEAGHER
STEVEN G. DAVIS
DEIRDRE E. SANDERS

SANDRA A. BROCKMAN-LEE
CHRISTOPHER P. CARROLL
R. JAMES COE
CHRISTINE M. DOE
COLIN C. DURHAM
CAROL A. EONER
ERIK L. ENGE
GIOVANNA FISHENDEN
CAROLINE M. FLEMING
TODD A. GRETY
HELEN LEE
JOSEPH M. MATAIA
MARY K. MURRAY
KEVIN T. SHAUGHNESSY
MARK B. SOLOMON
TROY T. SVIHL
RALPH TREMENTOZZI
DARRELL L. WONG

OF COUNSEL
ELIZABETH W. MATA

PATENT AGENTS
SUSAN M. ABDELHKA
ALEXANDER AKERHEZER
KHAIO ANDERSON
JESSE A. FECKER
LUCY LUBASHEV
PAMELA A. TORPEY
KAREN J. TOWNSEND
ROBERT H. UNDERWOOD

TECHNOLOGY SPECIALISTS
KAMILAH ALEXANDER
PAUL G. ALLOWAY
SUSAN C. KELLY
DOOYONG SHIM LIM
VIVIEN J. TANNOUS-MAGIN
MICHAEL M. YAMAUCHI

MICHAEL KOWALSKI
ADMINISTRATIVE DIRECTOR

BARBARA J. FOUGHT
ADMINISTRATOR OF
PATENT AND
TRADEMARK PRACTICE

August 5, 2004

Via Facsimile
Ms. Lynne Black
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313
Fax No. 703-305-0317

Re: Method and System for Finding Similar Records in Mixed Free-Text and Structured Data
Application Serial No. 10/091,932
Filing Date: March 6, 2002
Our Docket No. 2471.2001-001

Dear Examiner Black:

Per your request, enclosed is a copy of Reference AR, "Foundations of Statistical Natural Language Processing" by C. Manning and H. Schutze, on pages 539-44, previously cited in the information disclosure statement dated August 2, 2002.

Applicants are submitting one further publication by Eric Bloedorn as requested in a Supplemental Information Disclosure Statement. The earliest publication date of this article was March 28-29, 2000.

Please do not hesitate to contact me if you have further questions.

Very truly yours,

David J. Thibodeau, Jr.

@PFDesktop\::ODMA/MHODMA/HBSR05;jManage;491486;1

the user any information that is not already contained in d_1. Clearly, a better design is to show only one of the set of identical documents, but that violates the PRP.

Another simplification made by the PRP is to break up a complex information need into a number of queries which are each optimized in isolation. In practice, a document can be highly relevant to the complex information need as a whole even if it is not the optimal one for an intermediate step. An example here is an information need that the user initially expresses using ambiguous words, for example, the query *Jaguar* to search for information on the animal (as opposed to the car). The optimal response to this query may be the presentation of documents that make the user aware of the ambiguity and permit disambiguation of the query. In contrast, the PRP would mandate the presentation of documents that are highly relevant to either the car or the animal.

A third important caveat is that the probability of relevance is only estimated. Given the many simplifying assumptions we make in designing probabilistic models for IR, we cannot completely trust the probability estimates. One aspect of this problem is that the *variance* of the estimate of probability of relevance may be an important piece of evidence in some retrieval contexts. For example, a user may prefer a document that we are certain is probably relevant (low variance of probability estimate) to one whose estimated probability of relevance is higher, but that also has a higher variance of the estimate.

The Vector Space Model

The *vector space model* is one of the most widely used models for ad-hoc retrieval, mainly because of its conceptual simplicity and the appeal of the underlying metaphor of using spatial proximity for semantic proximity. Documents and queries are represented in a high-dimensional space, in which each dimension of the space corresponds to a word in the document collection. The most relevant documents for a query are expected to be those represented by the vectors closest to the query, that is, documents that use similar words to the query. Rather than considering the magnitude of the vectors, closeness is often calculated by just looking at angles and choosing documents that enclose the smallest angle with the query vector.

In figure 15.3, we show a vector space with two dimensions, corre-

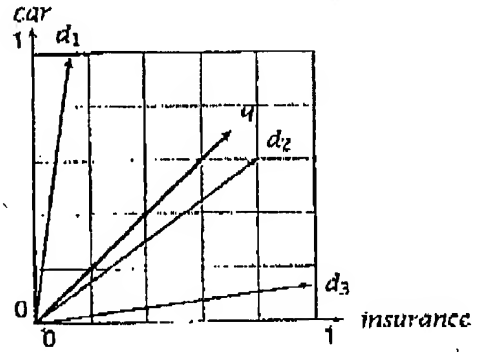

Figure 15.3 A vector space with two dimensions. The two dimensions correspond to the terms *car* and *insurance*. One query and three documents are represented in the space.

sponding to the words *car* and *insurance*. The entities represented in the space are the query q represented by the vector (0.71, 0.71), and three documents d_1, d_2, and d_3 with the following coordinates: (0.13, 0.99), (0.8, 0.6), and (0.99, 0.13). The coordinates or *term weights* are derived from occurrence counts as we will see below. For example, *insurance* may have only a passing reference in d_1 while there are several occurrences of *car* - hence the low weight for *insurance* and the high weight for *car*. In the context of information retrieval, the word *term* is used for both words and phrases. We say *term weights* rather than *word weights* because dimensions in the vector space model can correspond to phrases as well as words.)

In the figure, document d_2 has the smallest angle with q, so it will be the top-ranked document in response to the query *car insurance*. This is because both 'concepts' (*car* and *insurance*) are salient in d_2 and therefore have high weights. The other two documents also mention both terms, but in each case one of them is not a centrally important term in the document.

Vector similarity

To do retrieval in the vector space model, documents are ranked accord-

normalized correlation coefficient. We introduced the cosine as a measure of vector similarity in section 8.5.1 and repeat its definition here:

$$\cos(\vec{q},\vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} d_i^2}}$$

where $\vec{q}$ and $\vec{d}$ are n-dimensional vectors in a real-valued space, the space of all terms in the case of the vector space model. We compute how well the occurrence of term i (measured by q_i and d_i) correlates in query and document and then divide by the Euclidean length of the two vectors to scale for the magnitude of the individual q_i and d_i.

Recall also from section 8.5.1 that cosine and Euclidean distance give rise to the same ranking for normalized vectors:

$$\begin{aligned}(|\vec{x}-\vec{y}|)^2 &= \sum_{i=1}^{n}(x_i-y_i)^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \\ &= 1 - 2\sum_{i=1}^{n} x_i y_i + 1 \\ &= 2\left(1 - \sum_{i=1}^{n} x_i y_i\right)\end{aligned}$$

So for a particular query $\vec{q}$ and any two documents $\vec{d_1}$ and $\vec{d_2}$ we have:

$$\cos(\vec{q},\vec{d_1}) > \cos(\vec{q},\vec{d_2}) \Leftrightarrow |\vec{q}-\vec{d_1}| < |\vec{q}-\vec{d_2}|$$

which implies that the rankings are the same. (We again assume normalized vectors here.)

If the vectors are normalized, we can compute the cosine as a simple dot product. Normalization is generally seen as a good thing – otherwise longer vectors (corresponding to longer documents) would have an unfair advantage and get ranked higher than shorter ones. (We leave it as an exercise to show that the vectors in figure 15.3 are normalized, that is, $\sqrt{\sum_i \vec{d_i}^2} = 1$.)

Term weighting

We now turn to the question of how to weight words in the vector space model. One could just use the count of a word in a document as its term

	Symbol	Definition
ency	$tf_{i,j}$	number of occurrences of w_i in d_j
frequency	df_i	number of documents in the collection that w_i occurs in
requency	cf_i	total number of occurrences of w_i in the collection

Table 15.3 Three quantities that are commonly used in term weighting in information retrieval.

Word	Collection Frequency	Document Frequency
insurance	10440	3997
try	10422	8760

Table 15.4 Term and document frequencies of two words in an example corpus.

weight, but there are more effective methods of term weighting. The basic information used in term weighting is *term frequency*, *document frequency*, and sometimes *collection frequency* as defined in table 15.3. Note that $df_i \leq cf_i$ and that $\sum_j tf_{i,j} = cf_i$. It is also important to note that document frequency and collection frequency can only be used if there is a collection. This assumption is not always true, for example if collections are created dynamically by selecting several databases from a large set (as may be the case on one of the large on-line information services), and joining them into a temporary collection.

The information that is captured by term frequency is how salient a word is within a given document. The higher the term frequency (the more often the word occurs) the more likely it is that the word is a good description of the content of the document. Term frequency is usually dampened by a function like $f(tf) = \sqrt{tf}$ or $f(tf) = 1 + \log(tf), tf > 0$ because more occurrences of a word indicate higher importance, but not as much relative importance as the undampened count would suggest. For example, $\sqrt{3}$ or $1 + \log 3$ better reflect the importance of a word with three occurrences than the count 3 itself. The document is somewhat more important than a document with one occurrence, but not three times as important.

The second quantity, document frequency, can be interpreted as an indicator of informativeness. A semantically focussed word will often occur several times in a document if it occurs at all. Semantically unfocussed words are spread out homogeneously over all documents. An example

from a corpus of *New York Times* articles is the words *insurance* and *try* in table 15.4. The two words have about the same collection frequency, the total number of occurrences in the document collection. But *insurance* occurs in only half as many documents as *try*. This is because the word *try* can be used when talking about almost any topic since one can *try* to do something in any context. In contrast, *insurance* refers to a narrowly defined concept that is only relevant to a small set of topics. Another property of semantically focussed words is that, if they come up once in a document, they often occur several times. *Insurance* occurs about three times per document, averaged over documents it occurs in at least once. This is simply due to the fact that most articles about health insurance, car insurance or similar topics will refer multiple times to the concept of insurance.

One way to combine a word's term frequency $\mathrm{tf}_{t,d}$ and document frequency df_t into a single weight is as follows:

$$\text{weight}(t,d) = \begin{cases} (1+\log(\mathrm{tf}_{t,d}))\log\frac{N}{\mathrm{df}_t} & \text{if } \mathrm{tf}_{t,d} \geq 1 \\ 0 & \text{if } \mathrm{tf}_{t,d} = 0 \end{cases} \quad (15.5)$$

where N is the total number of documents. The first clause applies for words occurring in the document, whereas for words that do not appear ($\mathrm{tf}_{t,d} = 0$), we set $\text{weight}(t,d) = 0$.

Document frequency is also scaled logarithmically. The formula $\log\frac{N}{\mathrm{df}_t} = \log N - \log \mathrm{df}_t$ gives full weight to words that occur in 1 document ($\log N - \log \mathrm{df}_t = \log N - \log 1 = \log N$). A word that occurred in all documents would get zero weight ($\log N - \log \mathrm{df}_t = \log N - \log N = 0$).

This form of document frequency weighting is often called *inverse document frequency* or idf weighting. More generally, the weighting scheme in (15.5) is an example of a larger family of so-called *tf.idf* weighting schemes. Each such scheme can be characterized by its term occurrence weighting, its document frequency weighting and its normalization. In one description scheme, we assign a letter code to each component of the tf.idf scheme. The scheme in (15.5) can then be described as "ltn" for logarithmic occurrence count weighting (l), logarithmic document frequency weighting (t), and no normalization (n). Other weighting possibilities are listed in table 15.5. For example, "ann" is augmented term occurrence weighting, no document frequency weighting and no normalization. We refer to vector length normalization as cosine normalization because the inner product between two length-normalized vectors (the query-document similarity measure used in the vector space model) is